Finding your way through the woods with GrETEL

Liesbeth Augustinus, Vincent Vandeghinste,
Ineke Schuurman, and Frank Van Eynde
CCL, KU Leuven
{liesbeth,vincent,ineke,frank}@ccl.kuleuven.be

GrETEL (**Gr**eedy **E**xtraction of **T**rees for **E**mperical **L**inguistics [Augustinus et al., 2012] is a linguistic search engine enabling linguists to consult a syntactically annotated corpus (or treebank) in a very easy way.[1] As a starting point for searching the treebank, GrETEL takes a natural language example (e.g. a short sentence) instead of a complex search instruction. Therefore, limited or no knowledge about tree representations and formal query languages is needed. By allowing linguists to search for constructions which are similar to the example they provide, GrETEL allows user-friendly access to resources without spending time on technical details.

Making use of a treebank instead of a 'flat' corpus is especially recommended when looking for possibly discontinuous constructions, e.g. verbs with a fixed preposition such as *kijken naar* 'look at' in (1). In a flat corpus, unwanted constructions as (2) will pop up in the results while looking for sentences with a combination of *kijken* and *naar*, but in a treebank, they will not.

(1) Hij **keek** met een bang hartje **naar** de heks.
He looked with a scared little heart at the witch
'He was looking at the witch with a heavy heart.'

(2) Hij keek op zijn horloge terwijl hij naar de bushalte stapte.
He looked at his watch while he to the bus stop walks
'He looked at his watch while he was walking to the bus stop.'

If one would query the treebank for constructions as (1) with a formal query language, one would have to construct an expression as in (3). But for GrETEL, the Dutch sentence in (1) is sufficient.

(3)
```
//node[@cat="smain" and node[@rel="hd" and @pos="verb" and
@root="kijk"] and node[@rel="ld" and @cat="pp" and
node[@rel="hd" and @pos="prep" and @root="naar"]]]
```

[1] http://nederbooms.ccl.kuleuven.be/eng/gretel

In our presentation we will show how several types of discontinuous constructions can easily be extracted from the Dutch LASSY[2] [van Noord et al., 2013] and CGN[3] [Hoekstra et al., 2003] treebanks in order to use them for research in linguistics.

Keywords: treebank, querying, search tool

References

L. Augustinus, V. Vandeghinste, and F. Van Eynde. Example-Based Treebank Querying. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, 2012.

H. Hoekstra, M. Moortgat, B. Renmans, M. Schouppe, I. Schuurman, and T. van der Wouden. *CGN Syntactische Annotatie*, 2003. 77p.

G. van Noord, G. Bouma, F. Van Eynde, D. de Kok, J. van der Linde, I. Schuurman, E. Tjong Kim Sang, and V. Vandeghinste. Large Scale Syntactic Annotation of Written Dutch: Lassy. In *Essential Speech and Language Technology for Dutch: resources, tools and applications*. Springer, 2013.

[2] `http://odur.let.rug.nl/~vannoord/Lassy`

[3] `http://tst-centrale.org/images/stories/producten/documentatie/cgn_website/doc_English/start.htm`